DOCUMENT RESUME

ED 273 659 TM 860 508

AUTHOR Rowley, Glenn L.
TITLE Application of Generalizability Theory to Observational Studies: Limitations.
PUB DATE Apr 86
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Classroom Observation Techniques; *Classroom Research; Correlation; Elementary Education; *Error of Measurement; *Generalizability Theory; *Interrater Reliability; Learning Disabilities; Mathematical Models; Observation; *Regression (Statistics)

ABSTRACT

Classroom researchers are frequently urged to provide evidence of the reliability of their data. In the case of observational data, three approaches to this have emerged: observer agreement, generalizability theory, and measurement error. Generalizability theory provides the most powerful approach given an adequate data collection design, but observational data provide an opportunity to obtain direct measures of the magnitude of measurement error in a way that is not possible with other kinds of data used in educational research. Two methods of assessing measurement error are demonstrated--by direct measurement and by regression estimates based on mean score. An example, using 105 primary children in self-contained classrooms, investigates the influence of classroom processes on the reading performance of learning disabled students. The argument is made that there is no one correct way of expressing the adequacy of data for all purposes; instead one must turn attention to the specific indicators of data quality that are most relevant to the intended use of the data. A table outlines indicators of adequacy of data. (Author/JAZ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED273659

APPLICATION OF GENERALIZABILITY THEORY TO OBSERVATIONAL STUDIES: LIMITATIONS

Glenn L. Rowley
Monash University, Australia

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. L. Rowley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 860 508

Paper presented as part of a symposium "New Developments in Generalizability Theory: International Perspectives", at the 1985 Annual Meeting of the American Educational Research Association. San Francisco, April 1986.

APPLICATION OF GENERALIZABILITY THEORY TO OBSERVATIONAL STUDIES: LIMITATIONS

Glenn L. Rowley

Monash University

ABSTRACT

Classroom researchers are frequently urged to provide evidence of the reliability of their data. In the case of observational data, three approaches to this have emerged: observer agreement, generalizability theory, and measurement error. Generalizability theory provides the most powerful approach given an adequate data collection design, but observational data provide an opportunity to obtain direct measures of the magnitude of measurement error in a way that is not possible with other kinds of data used in educational research. Two methods of assessing measurement error are demonstrated - by direct measurement and by regression estimates based on mean score. The argument is made that there is no one correct way of expressing the adequacy of data for all purposes; instead we must turn our attention to the specific indicators of data quality that are most relevant to the intended use of the data. A framework is outlined for doing this.

Guidelines for users and developers of classroom observation systems (Herbert & Attridge, 1975) have emphasized the need to provide data pertaining to reliability, along with "a discussion of which reliability measures were selected, and why (p. 14)," but they have provided little in the way of guidance as to how this can be done.

Three approaches to the problem have emerged. The first, and still the most widely used, is that which focusses attention on agreement between observers, typically reporting a percentage figure, and less frequently a between-observer correlation coefficient. More recent work has been conducted within the framework of generalizability theory, focussing on the variability associated with observers, occasions, situations, and individuals. And finally, some recent work has approached the problem from a sampling perspective, which allows for the direct estimation of measurement error.

Examination of the use of these methods in reported research, however, reveals that the type of reliability data reported is determined in large part by "journal culture" - for example, in the Journal of Applied Behavior Analysis it is customary to report observer agreement coefficients, while in the American Educational Research Journal, the presentation of generalizability information is more acceptable. In this paper it is argued that the choice of information presented should be determined, not by the journal culture, but by the use which is to be made of the data.

APPROACHES TO THE RELIABILITY OF OBSERVATIONAL MEASURES

Observer Agreement

The tradition of estimating reliability via observer agreement is long and well-established. In fact, there exists a whole range of procedures which fall under this general heading. The procedures differ from one another with respect to the kind of data upon which agreement is computed (individual categories versus more complex scales), with respect to design features (who are the observers and how are the observations arranged?), and with respect to computational procedures -- in particular, whether one reports raw percentages of agreement, the percentage by which agreement exceeds some change expectation, or correlation coefficients.

Observer agreement methods have been adequately reviewed by Frick & Semmel (1978) and by Mitchell (1979). It is widely agreed that high levels of observer agreement are necessary but not sufficient evidence

that measures are reliable. This is because observer agreement indices do not count as error the notorious instability of human behaviour from one occasion to the next. A high coefficient of observer agreement can provide some confidence that one particular occasion has been accurately observed, but no basis for generalizing about the behaviour of that person or group of persons over time.

Generalizability

Although the early work of Medley and Mitzel (1963) was not phrased in its language, their ideas were clearly the ideas central to the development of generalizability theory of Cronbach, Gleser, Nanda and Rajaratnam (1972). For example, Medley and Mitzel were careful to distinguish between the reliability coefficient, and the different, but associated concepts of observer agreement and stability. Readers familiar with generalizability theory will recognize these concepts as three different generalizability coefficients, corresponding to three different universes of generalization, in the language provided by Cronbach et al (1972). The factorial design used for reliability estimation and the analysis-of-variance procedures employed similarly anticipated the later work in generalizability theory.

The link with generalizability theory was made explicit by McGaw, Wardrop and Bunda (1972) who, in addition to basing their approach on the concepts of generalizability theory, insisted that variation among situations is a legitimate source of systematic variance which ought to be quantified. The effect of this consideration is to add another factor (occasions within situations) to an already complex design. Rowley (1976) demonstrated the use of a simpler design which enabled the classroom researcher to estimate reliability for data collected in the normal course of a research study, rather than a study particularly designed to estimate variance components and hence generalizability coefficients. In the terminology used by Cronbach et al. (1972), reliability can be assessed routinely as part of a decision (D-) study, without requiring the researcher to conduct a separate generalizability (G-) study.

It is widely agreed that the questions we commonly ask about reliability are better seen within the general framework of generalizability theory (Cronbach et al, 1972). Questions about "reliability" have thus been equated with questions about the contribution of relevant facets of observation to variation in observed behavior. The solution that has been arrived at is elegant, satisfying, and (given an

appropriate data-collection design and sufficient data) widely usable. It answers a broad range of questions, not just one. It is widely accepted, but less widely used in classroom research. Perhaps this is because it has provided answers to questions that classroom researchers had not (in general) been asking.

Measurement Error

A third approach is possible, and this involves the direct assessment of the magnitude of measurement errors. Because classroom observation measures do more than just provide a rank-ordering of the subjects of measurement, the usefulness of the reliability coefficient is limited. The problem is the same one which Popham and Husek (1969), along with many others have documented with respect to criterion-referrenced measurement. The reliability coefficient describes how effectively a measure discriminates among the subjects of observation, not how precisely it measures them. Precision of measurement is another matter, and one which is readily addressed in the context of classroom observation.

Cooley and Mao (1981) have shown how the magnitude of measurement error can be assessed by approaching the problem from a sampling perspective. In the simplest case, suppose that a person's behaviour is observed on 100 distinct moments, and that one particular behaviour X was observed for eight of those 100 observations. Then we could estimate the probability of going into that classroom and observing that particular behaviour to be $8 \div 100$, or .08. If we are willing to regard the occasions observed as a random sample from a universe of occasions, we may estimate the expected score over that universe to be .08 with a standard error given by $\sqrt{.08\ (1-.08)/100}$. With some degree of approximation, the standard error can be translated into a confidence interval. It should be noted that this method requires fairly strong assumptions concerning the distribution of observations over time, and the occurrence of the behaviour being observed. The means of assessing measurement error that is outlined below does not require these assumptions, and takes account of variability across occasions.

ESTIMATION OF ERROR FROM VARIANCE OVER OCCASIONS

The nature of classroom observation data typically provide us with an opportunity to calculate standard errors directly -- an opportunity which is almost unique outside the physical sciences. Nobody is better-placed

than the classroom researcher to estimate standard errors of measurement; for each subject of measurement, there will usually be several equivalent measures of the same behaviour. The standard deviation of the individual measures is certainly of interest and can be reported directly. If, as is likely, the measurement of interest is the mean score for each person, then the standard error of that mean score can and should be reported. And (perhaps this is almost the only situation in educational research where we are afforded this luxury) we can, if we wish, obtain a separate standard error for each person observed.

In practice, a standard error, SE(i) for an individual i can be computed by obtaining a measure X(i,j) from that person for each period j of observation, and then using the formula for the standard error of the mean; in this case $SE(i) = S(i)/\sqrt{n}$ where S(i) = standard deviation of the measures X(i,j) (j = 1, ..., n), and n = number of periods of observation for individual i.

Because n would typically be small, the estimates of SE(i) would be imprecise, and it would in most cases be preferable to average them across individuals, and report a single standard error for the group, just as is effectively done with test scores. However, it is possible, even likely, that the size of the standard error for an individual would be systematically related to mean score (or to other variables of interest), in which case greater precision could be obtained by regressing individual standard errors against these variables, and using regression estimates for individuals, which would be less subject to sampling fluctuations.

An Example

But how important is it that estimates of precision should take account of such differences? Some exploratory work has been done with data gathered in the course of a larger study, investigating the influence of classroom processes on the reading performance of learning disabled students (Leinhardt, Zigmond & Cooley, 1981), and the results suggest that it may be very important.[1]

The data are described in detail by Leinhardt et al. (1981). The essentials are as follows: One hundred and five children in 11 primary grade self-contained classrooms in Pittsburgh were observed over a 20-week

---

1. The author is grateful to Gaea Leinhardt and William Cooley, of the Learning Research and Development Center, Pittsburgh, who made available the data used in this study.

period. For each classroom, approximately 30 one-hour observation periods were obtained (approximately 20 in mornings and 10 in afternoons). The instrument used was the Student-Level Observation of Beginning Reading (Leinhardt & Seewald, 1981) but in this study, only two variables have been considered. For each observation period, we have recorded the number of times (out of 12) that the student was observed to be engaged in -

Direct Reading: oral silent reading of letters, words, sentences and paragraphs, and
Indirect Reading: activities which were assumed to be related to reading, but in which the student was not engaged in responses to print in the "normal" way (e.g. story discussion, listening and writing).

Morning and afternoon data were kept separate: we therefore have two occasions and two variables, which in investigating precision have been treated as separate sets of data (in a generalizability study, morning and afternoon would constitute of "situations" facet). The data take the form presented in Table 1.

Insert Table 1 about here

Since there was a total of 2325 observations made on the 105 pupils, Table 1 includes only a sample of the data: the first three children and the last two. In this case there are enough data from each pupil to provide reasonable estimates of the mean scores, and of their standard errors. In the table, both are presented in the metric of the observations (number of times observed to be engaged in direct reading out of 12 opportunites); these could be translated into proportions or percentages. In the latter case, the means would be 14.4%, 14.8%, 23.1%, ... , 5.0%, 8.3%, and the standard errors 3.1%, 3.4%, 3.5% ... , 3.1%, and 2.4%.

The particular data selected arbitrarily for presentation in Table 1 suggest (wrongly, as it transpires) that a single estimate of precision (the standard error figure) would suffice for all individuals, as would be necessary if errors were computed from standard reliability or generalizability analyses. For data such as these, it would be surprising if this were so. If the occurrences were random and infrequent, the

observed frequencies would follow a Poisson distribution, for which the average frequency of occurrence would be linearly related to the standard deviation - and hence, for a fixed number of observations, to the standard error.

For these data, the occurrences are, of course, not random, but they are infrequent. Figure 1 despicts the relation between mean and standard deviation, and the linear relationship is obvious (the correlation is 0.86). Students who display the behavior being observed more frequently are also more variable in the extent to which they display it, and are therefore measured with less precision. This suggests more strongly that a single estimate of precision will not suffice for all individuals.

Insert Figure 1 about here

The linearity of the regression suggests also that better estimates of the standard error for an individual may be obtained by taking group data into account, just as we do in computing regression estimates of true score. For these data, the regression equation is

$$\text{est. S.D.} = .70 + .60\ (\text{Mean})$$

and data on precision could be adequately described by a table such as Table 2. From the remaining three sets of data, similar relationships

---

Insert Table 2 about here

---

were observed (the correlation coefficients were 0.86, 0.87, 0.46 and 0.79). The remaining entries in Table 2 were obtained using regression equations in the manner described above. Another common feature of the four relationships was heteroscedasticity, so that the estimates of standard error are seen to be more accurate for low scores than for high scores.

HOW IS DATA QUALITY BEST DESCRIBED?

For years researchers have posed this question with respect to observational data. Our basic mistake has been to assume that the question has an answer. Instead of asking, "How should we best assess reliability?", and expecting to find a single answer, it would have been more appropriate to ask, "What uses will be made of the observational data?", and then "What measurement properties will the data need to have if they are to prove adequate for these uses?"

In order to undertake this task, we need to begin with some categorization of the uses to which classroom observation data are put. This will be organized around a framework which derives from research studies conducted by Gage and some of his associates. While not exhaustive, this framework does include the major uses of observational data in research, and it provides a conceptual structure within which we can look at the measurement properties which are likely to be important for the various uses of the data.

Gage and his co-workers (e.g. Gage & Coladarci, 1980) have been engaged over recent years in work which attempts to complete the "descriptive-correlational-experimental loop" described by Rosenshine and Furst (1973) as the only way to obtain firm evidence of causal links between processes and outcomes in teaching. Classroom observation is intimately involved in each of the four steps in their research strategy:

1. Descriptive studies: development of observation instruments, identification of potentially important classroom variables.
2. Correlational studies: identification of classroom process variables which correlate with desired learning outcomes, and which therefore have potential importance in experimental studies.
3. Training of teachers in clusters of potentially important behaviors.
4. Experimental studies: in which teachers so trained are matched with a control group, and pupil outcomes compared. Observation at this stage is focused on the degree of implementation of the training, and possibly also on pupil behaviors.

Sections to follow will give consideration to some of the uses to which observational data are put in each step, and to some of the reliability-related issues with can arise.

Descriptive Studies

In the development of an instrument for classroom observation, the major issues to be addressed are the following:

1. Does the instrument focus the observer's attention on the most important behaviors for the purpose in hand?
2. Are the instructions to the observer (both with regard to what to observe, and how to code observations) clear and unambiguous?

The first question relates to what might broadly be referred to as validity, and has both judgmental and empirical aspects. The second question falls within the domain which is frequently identified as reliability. If two observers categorize a particular behavior in the same way, this would indicate, without proving, that the instructions to the observers are sufficiently clear, and the behaviors identifiable. Thus if the instrument meets the minimal criterion of observer agreement, based on individual observations, it is agreed to be adequate. Observer agreement is a legitimate concern at the stage in which observational procedures are being developed, tried out, and modified. Satisfactory observer agreement establishes that the instrument is usable, but not that the measures derived from it are useful in any particular research context. And it does not, of course, establish reliability.

In a purely descriptive study (e.g., one in which the aim is restricted to describing the activities of one or more teachers over a fixed time-period), the only concern is accuracy of observation -- do the recorded observations provide an accurate record of what those teachers did while they were being observed? High observer agreement does not prove this, by any means, since two observers can agree and still be wrong. But is does provide supportive evidence, in that the data had the opportunity to disprove the accuracy of the observations, and they did not. Thus observer agreement is a relevant concern in what we have described as a purely descriptive study.

But few studies are purely descriptive in the sense described above; inference and generalization almost always play a part. It is difficult to conceive of classroom research in which no time period other than the period observed was of interest. Even if the researcher does not want to generalize across teachers, or classrooms, almost invariably the intention will be to generalize across time -- to treat any measure as though it describes a characteristic of that person, or of that person's

behavior. The adequacy of that generalization is clearly a matter for concern, and can be addressed either through generalizability coefficients, or through the direct estimation of the magnitudes of measurement errors. Generalizability coefficients would be most useful where the focus of the study is on relative standings; measurement error would be favoured where the aim is simply to provide accurate description.

Correlational Studies

Typically, correlational studies have been used in attempt to identify teacher behaviors which relate to valued outcomes. Sometimes simple correlations are reported; on other occasions, stronger inferences may be sought through the use of causal modelling. Clustering techniques are frequently used to group behaviors into manageable composites. Even though the statistics may be complex, we should not let ourselves forget that they are based on correlations, and that the inferences which may be drawn require a degree of caution.

Since Spearman (1904), we have known that the reliability of a measure provides a limit to the extent to which that measure can correlate with any other measure. In any correlational study, therefore, reliability is a necessary property. No matter what the attribute, if it is measured with low reliability, it will correlate poorly with any criterion, even in circumstances where the attribute has a strong causal influence on the criterion. It is easy to forget that a low correlation between two variables does not establish lack of causality; low correlations can, and probably often do, arise simply from the use of unreliable measures.

Reliability as assessed by a specific generalizability coefficient is the most important pre-requisite for measures which are to be used in correlational studies. This interpretation of correlational studies includes any research in which inferences are drawn from the observed associations among variables, whether or not the associations are formally quantified using correlation coefficients. Informal hypothesizing in an anthropological study, and the economist's predictions of an upward move in interest rates from a set of economic indicators would be subject to the same limitations. The limitations, though, may not be readily quantifiable in these cases because no means are known for investigating the measurement properties of the observations involved.

Training Teachers

Observations can be used to monitor the effectiveness of training programs. The behavior of teachers before and after training can be observed and crucial behaviors measured, just as tests are administered to pupils pre-and post-instruction. The measurement considerations are identical, and frequently misunderstood.

Reliability is not the property of a test which determines its capacity to distinguish between instructed and unistructed groups, or between the same group pre- and post-instuction. The most reliable test is the one which most effectively distinguishes among individuals, before or after instruction. This requires a large variance on each occasion. But the ideal situation for identifying change is one in which there is a small variance on each occasion, and a large difference between the means of the two occasions. The most reliable measure is not necessarily the one which is most sensitive to changes in individuals. Sensitivity to change implies between-group variability; reliability implies within-group variability.

Where classroom observation is used to monitor the effects of training on behavior, the standard error is again a useful statistic. But perhaps more directly relevant is the standard error of the group mean, since from it we can simply deduce the magnitude of the changes we will be able to detect. If we find, e.g., that the standard error of the group mean has magnitude comparable to the change we can reasonably anticipate, we clearly need to obtain more precise measurements, perhaps by scheduling more observation. The point is that attention to the standard error of the mean will provide a better indication of whether sufficient precision has been achieved than will attention to a reliability coefficient.

Experimental Studies

Observation is not directly involved in the experimental phase of Gage's research strategy, except in so far as it is used to monitor the implementation of the treatments by experimental and control groups, and to establish that the training program really did influence the teachers' behavior. The considerations are similar to those in the previous section, and again a reliability coefficient does not tell us directly what we want to know. A reliability coefficient provides a comparison between measurement error and within-group variance; more relevant to our concerns would be a comparison between measurement error and between-group variance. Prior to the running of the experiment, this is not

possible unless there are data available upon which estimates of between-group differences can be based. Knowledge of the standard error of measurement, however, enables the exploration of the adequacy of the measures for detecting differences of various magnitudes.

CONCLUSION

Quality of data remains a most important consideration in classroom research, as it must in all research. But data have many qualities, not just one. In the past we have been too ready to believe that all important measurement qualities were enshrined in a single coefficient. There are uses of classroom observation data for which high generalizability coefficients are vitally important, and there are uses for which other qualities are more important. Consideration of data quality cannot be separated from the uses which are going to be made from the data.

Just as issues in the general area known as test validity can be subsumed under the general question, "What evidence supports the inferences being made from the test scores?", so issues in the area generally thought of as reliability can and should be subsumed under a similar question: "Are the measurement qualities possessed by the data adequate for the purpose for which they are being used?" Table 2 provides a categorization of some of the inferences drawn from observational data, along with an assessment of the measurement properties needed if the data are to be capable of supporting these inferences.

INSERT TABLE 2 ABOUT HERE

To summarize the conclusions which have been reached: in correlational studies we need to be concerned about the reliability of observational measures, but in descriptive studies, and in detecting changes in behaviors, standard errors provide more directly useful information. Furthermore the standard error of measurement for a group, and even for individuals, can be estimated in a reasonably straightforward manner, given the nature of observational data. The reporting of this statistic with observational data should be encouraged. Measures of observer agreement establish the usability of an observational instrument, and should be encouraged during the instrument-development stage. Because they do not admit variation over occasions as a source of error, they do

not tell us very much about the potential usefulness of the instrument in research. Generalizability studies are useful because they allow us to estimate some aspects of data quality even when different data-collection designs are used.

In general, we should recognize that observational data have many uses, and that a reliability coefficient provides evidence of the suitability of the data for only some of those uses. There is no single indicator which demonstrates the adequacy of data for all possible uses; instead we must turn our attention to the specific indicators of data quality which are most relevant to the intended use of the data.

REFERENCES

Cooley, W.W., & Mao, B.J. (1981). The sample of classroom time observed. Journal of Classroom Interaction, 17(1), 31-36.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). The Dependability of Behavioral Measurements: Generalizability of scores and profiles. New York: Wiley.

Frick, T. and Semmel, M.I. (1978). Observer agreement and reliabilities of observational measures. Review of Educational Research, 48, 157-184.

Gage, N.L. and Coladarci, T. (1980). Replication of an experiment with a research-based inservice teacher education program. Final report on Grant No. NIE-G-79-0014, Stanford University.

Herbert, J.D. and Attridge, C. (1975). A guide for developers and users of observational systems. American Educational Research Journal, 12, 1-20.

Leinhardt, G. & Seewald, A.M. (1981). Student-level observation of beginning reading. Journal of Educational Measurement, 18, 171-177.

Leinhardt, G., Zigmund, N. & Cooley, W.W. (1981). Reading instruction and its effects. American Educational Research Journal, 18, 343-361.

McGaw, B., Wardrop, J.L. & Bunda, M.A. (1972). Classroom observation schemes -- where are the errors? American Educational Research Journal, 9, 13-27.

Medley, D.M. & Mitzel, H. (1963). Measuring classroom behavior by systematic observation. In N.L. Gage (ed.) Handbook of Research on Teaching. Chicago: Rand-McNally.

Mitchell, S.K. (1979). Interobserver agreement, reliability and generalizability of data collected in observational studies. Psychological Bulletin, 86, 376-390.

Popham, W.J., & Husek, T.R. (1969). Implications of criterion-referenced measurement. Journal of Educational Measurement, 6, 1-9.

Rosenshine, B. & Furst, N. (1973). The use of direct observation to study teaching. In R.M.W. Travers (ed.) Second Handbook of Research on Teaching. Chicago: Rand-McNally.

Rowley, G.L. (1976). The reliability of observational measures. American Educational Research Journal, 13, 51-59.

Spearman, C. (1904). The proof and measurement of association between two things. American Journal of Psychology, 15, 72-101.

TABLE 1

SAMPLE OF DATA FROM LEINHARDT ET AL (1981)
(DIRECT READING: AM ONLY)

Person ID	Direct Reading: No. of blocks out of 12 observed	No. of obsns.	Mean	SD	SE
1	0,2,0,5,0,3,1,1,2,1,0,1, 5,6,1,0,2,2,2,1,3,0	22	1.73	1.75	0.37
2	3,2,0,2,0,0,2,0,0,3,1,0, 4,5,4,1,4,1,2,2,0,3	22	1.77	1.60	0.34
3	4,2,0,0,0,0,1,3,5,2,2,0, 8,4,3,3,4,3,2,0,3,0	22	2.77	1.95	0.42
. . . . .					
104	0,0,0,0,2,0,0,0,0,0,7,0, 0,0,0,0,0,2,0,1	20	0.60	1.64	0.37
105	2,0,1,0,3,2,0,4,3,2,0,1, 0,0,0,0,0,0,0,2	20	1.00	1.30	0.29

TABLE 2

ESTIMATED INDIVIDUAL STANDARD ERRORS FOR DIFFERENT SCORES:
DATA FROM LEINHARDT ET AL. (1981)

Variable	Time	Number of Observations	Estimated Percent	Standard Error Percent
Direct	A.M.	20	10	2.6
			20	4.0
			30	5.3
			40	6.7
Direct	P.M.	10	10	3.9
			20	7.1
			30	10.4
			40	13.6
Indirect	A.M.	20	10	3.1
			20	3.7
			30	4.4
			40	5.0
Indirect	P.M.	10	10	3.8
			20	6.3
			30	8.8
			40	11.3

TABLE 3

INDICATORS OF ADEQUACY OF DATA

Inference to be made from Data	Indicator of adequacy of Data
Instrument is usable	Observer agreement
Accurate description of a single occasion	Observer agreement
Accurate description of a person's behavior over time	Standard errors of individual scores
Accurate description of variations in behavior among people	Reliability or generalizability coefficients
Portrayal of the relationship between behavior and other variables	Reliability coefficient
Detection of changed behavior as a result of an intervention	Standard error of the group mean
Relative contribution of various sources of variability	Generalizability coefficients

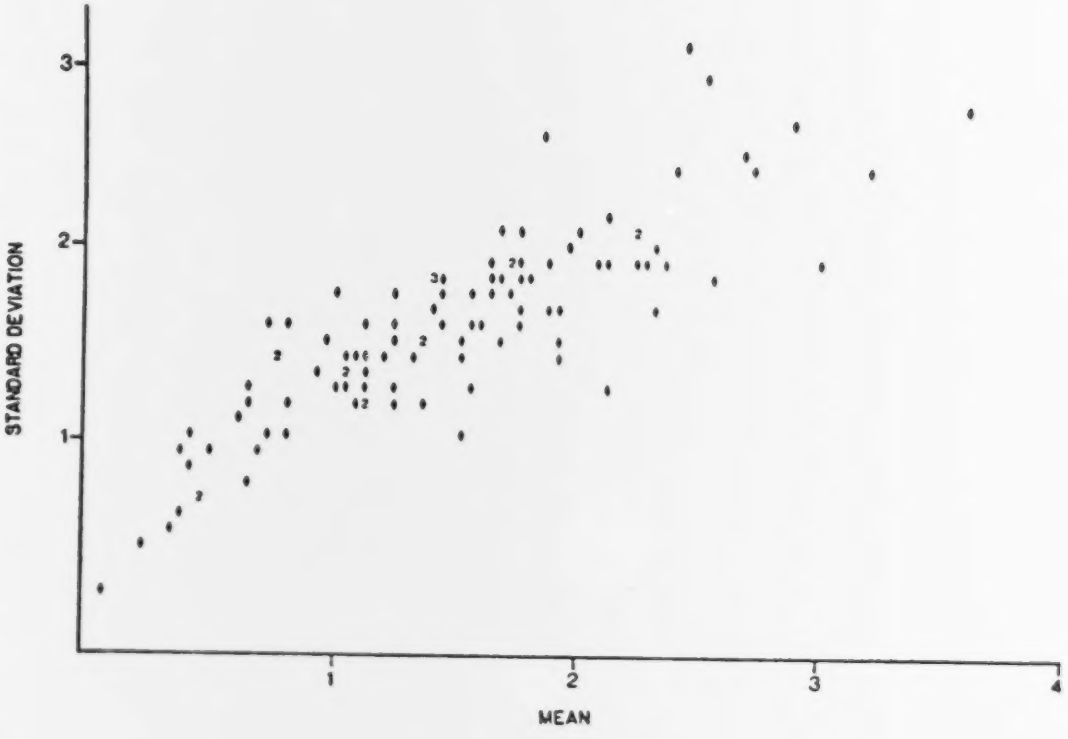

Figure 1

VARIABILITY VERSUS MAGNITUDE OF SCORE: A.M. DATA